Efficiently Optimizing over (Non-Convex) Cones via Approximate Projections

Michael B. Cohen micohen@mit.edu
Massachusetts Institute of Technology
Chinmay Hegde chinmay@iastate.edu
Iowa State University
Stefanie Jegelka stefje@mit.edu
Massachusetts Institute of Technology
Ludwig Schmidt ludwigs@mit.edu
Massachusetts Institute of Technology

Abstract

Constrained least squares is a ubiquitous optimization problem in machine learning, statistics, and signal processing. While projected gradient descent is usually an effective algorithm for solving constrained least squares at scale, the projection operator is often the computational bottleneck, especially for complicated constraints. To circumvent this limitation, we extend recent work on *approximate projections* to a significantly broader range of constrained least squares problems. Our new variant of projected gradient descent is able to utilize approximate projections for any condition number and any conic constraint set (including non-convex cones).

1 Introduction

Constrained optimization is a core element in machine learning, statistics, and signal processing. Over the past decade, these fields have developed a sophisticated understanding of various constraint sets, both from an optimization perspective and from a statistical point of view. The prototypical problem in this area is constrained least squares: given a data or measurement matrix $X \in \mathbb{R}^{n \times d}$, observations $y \in \mathbb{R}^n$, and a constraint set $C \subseteq \mathbb{R}^d$, the goal is to find a minizer of

$$\min_{\theta \in C} \|X\theta - y\|_2^2 \,. \tag{1}$$

For various constraint sets C, this optimization problem can represent important estimators such as (kernel) ridge regression, the Lasso, and nuclear norm minimization. The success of these methods leads to a fundamental question:

For what constraint sets can we solve least squares efficiently?

We make progress on this question by introducing a new version of projected gradient descent that utilizes *approximate projections* [8]. Approximate projections relax the usual notion of projecting onto a set by allowing a relative approximation error. Prior work has shown that approximate projections can be significantly faster than their exact counterparts for many natural constraint sets (see Table 1). In some settings such as graph sparsity, *exactly* projecting onto the constraint set is an NP-hard problem, while an *approximate* projection can still be computed in nearly-linear time [9].

Although the aforementioned papers establish significantly faster running times, their results are limited to the *RIP (Restricted Isometry Property)* setting that is common in the compressive sensing literature [7]. From an optimization point of view, this is a highly restrictive assumption as it requires the condition number of the objective function, i.e., the quadratic loss $\|X\theta - y\|_2^2$, to be close to 1. Moreover, prior results require the condition number to improve even more as the projections become more approximate [8, 10].

In this paper, we generalize the approximate projection framework along multiple directions:

- We generalize the definitions of approximate projections to *any* conic constraint set (even non-convex cones).

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

Constraint set	Best known time complexity of an **exact projection**	Best known time complexity of an **approximate projection**
Sparsity	$O(d)$	$O(d)$
Low-rank matrices	$O(d^{1.5})$	$\widetilde{O}(r \cdot d)$ [10]
Tree sparsity [5]	$O(d^2)$ [6]	$\widetilde{O}(d)$ [2]
Graph sparsity [11]	NP-hard [9]	$\widetilde{O}(d)$ [9]
Group sparsity [15]	NP-hard [3]	$\widetilde{O}(d)$ [12]

Table 1: The time complexity of exact and approximate projections for various constraint sets. The variable d denotes the size of the input (the dimension of the vector θ). For low-rank matrices, the stated time complexites are for matrices with dimension $\sqrt{d} \times \sqrt{d}$. To simplify the expressions, we omit logarithmic factors in the runnin times. In all cases beyond simple sparsity, approximate projections are significantly faster (assuming P $\neq$ NP, even by a super-polynomial amount).

- We introduce a new variant of projected gradient descent (2PHASE-PGD) that can be combined with approximate projections for *any* condition number and *any* conic constraint set.
- We prove that 2PHASE-PGD achieves optimal statistical guarantees (up to constant factors) in all relevant problem parameters, including the condition number.

Interestingly, our new algorithm 2PHASE-PGD combines ideas from both the Frank-Wolfe algorithm and projected gradient descent. The outer loop of 2PHASE-PGD is essentially projected gradient descent, and each iteration of the outer loop calls an inner subroutine that is very similar to the Frank-Wolfe algorithm. The combination of these two parts allows our algorithm to utilize approximate projections without restrictions on the constraint sets and condition numbers.

1.1 Related work

Due to the vast literature on constrained least squares estimators, we only mention the most closely related works here.

The combination of approximate projections that we utilize in our work was first introduced in the model-based compressive sensing setting [8], where the goal is to add additional sparsity-based constraints to compressive sensing algorithms [4]. The authors later generalized their approach to general union-of-subspace models that include low-rank matrices [10]. However, all of their algorithms rely on very well conditioned objective functions (the RIP setting). Our algorithm lifts this restriction and applies to any condition number (and any conic constraint set).

The paper [14] also considers projected gradient descent for general constraint sets that can be non-convex. However, the goals of the paper are somewhat different from ours: the authors focus on establishing sharp constants for isotropic measurement setups or data matrices, which corresponds to the well-conditioned regime (RIP). In contrast, our focus is on the regime where the condition number can be arbitrary. Moreover, our algorithm works with approximate projections.

Finally, the paper [13] also considers iterative thresholding methods (i.e., projected gradient descent) in the regime where the condition number can be arbitrary. However, the sample complexity does not match the optimal rates achieved by the Lasso estimator because the statistical rate has a quadratic dependence on the condition number. Our two-phase variant of projected gradient descent addresses this shortcoming. Moreover, our algorithm works with approximate projections and arbitrary conic constraint sets.

2 Results

We state our results in the common setting where we assume *restricted* strong convexity and smoothness [1]. In contrast to the classical (global) notions of strong convexity and smoothness, the restricted counterparts only hold over the constraint set (or a suitably relaxed version of the constraint set). This weakened assumption is crucial because the concentration phenomena in statistical settings are only sufficient to guarantee strong convexity and smoothness over a subset of the entire parameter space.

Definition 1 (Restricted smoothness)**.** *Let C be a cone. Then a differentiable convex function f has restricted smoothness L over C if, for every point $x \in \mathbb{R}^d$ and every vector $u \in C$,*

$$\langle \nabla f(x+u) - \nabla f(x), u \rangle \leq L\|u\|_2^2.$$

Note that for a quadratic function $\frac{1}{2}\|Ax-b\|_2^2$, this is equivalent to $\|Au\|_2^2 \leq L\|u\|_2^2$ for every $u \in C$.

Definition 2 (Restricted strong convexity)**.** *Let C be a cone. Then a differentiable convex function f has restricted strong convexity ℓ over C if, for every point $x \in \mathbb{R}^d$ and every vector $u \in C$,*

$$f(x+u) \geq f(x) + \langle \nabla f(x), u \rangle + \frac{\ell}{2}\|u\|_2^2.$$

For a quadratic function $\frac{1}{2}\|Ax-b\|_2^2$, this is equivalent to $\|Au\|_2^2 \geq \ell\|u\|_2^2$ for every $u \in C$.

2.1 Approximate projections

Next, we introduce our new definitions of "head" and "tail" approximations, which are two complementary notions of approximate projections. These relaxed projections are the only way our algorithm 2PHASE-PGD interfaces with the constraint sets.[1]

Definition 3 (Head approximation)**.** *Let $C^\star, C_\mathcal{H} \subseteq \mathbb{R}^d$ be two cones and let $c_\mathcal{H} \in \mathbb{R}$. Then an $(C^\star, C_\mathcal{H}, c_\mathcal{H})$-head approximation satisfies the following property. Given any vector $g \in \mathbb{R}^d$, the head approximation returns a unit vector $\theta \in C_\mathcal{H} \cap \mathbb{S}^{d-1}$ such that*

$$\langle g, \theta \rangle \;\geq\; c_\mathcal{H} \cdot \max_{\theta' \in C^\star \cap \mathbb{S}^{d-1}} \langle g, \theta' \rangle \;.$$

Definition 4 (Tail approximation)**.** *Let $C^\star, C_\mathcal{H} \subseteq \mathbb{R}^d$ be two cones and let $c_\mathcal{T} \in \mathbb{R}$. Then an $(C^\star, C_\mathcal{T}, c_\mathcal{T})$-tail approximation satisfies the following property. Given any vector $\theta^{in} \in \mathbb{R}^d$, the tail approximation returns a vector $\theta \in C_\mathcal{T}$ such that*

$$\|\theta^{in} - \theta\|_2 \;\leq\; c_\mathcal{T} \cdot \min_{\theta' \in C^\star} \|\theta^{in} - \theta'\|_2 \;.$$

In addition to a relaxed approximation guarantee ($c_\mathcal{H} < 1$ and $c_\mathcal{T} > 1$, respectively), the above definitions also allow for approximation in the output sets $C_\mathcal{H}$ and $C_\mathcal{T}$. As long as $C_\mathcal{H}$ and $C_\mathcal{T}$ are comparable to $C^\star$ (say sparsity $2 \cdot s$ instead of sparsity s), this relaxation affects the sample complexity only by constant factors, yet enables polynomially faster algorithms in cases such as graph sparsity [9].

2.2 Main results

Before we state our main result, we briefly introduce some notation. For two sets C_1 and C_2, we denote the Minkowski sum by $C_1 + C_2$. For an integer m, we write $m \times C$ for the m-wise Minkowski sum of a set with itself (i.e., $C + C + C + \cdots + C$ a total of m times).

Theorem 5 (2PHASE-PGD)**.** *There is an algorithm* 2PHASE-PGD *with the following properties. Assume that* 2PHASE-PGD *is given*

- *a $(C^\star, C_\mathcal{T}, c_\mathcal{T})$-tail approximation such that $\theta^\star \in C^\star$, and*
- *a $(C^\star - C_\mathcal{T}, C_\mathcal{H}, c_\mathcal{H})$-head approximation.*

Then let

$$k \;=\; \Theta\left(\frac{(1+c_\mathcal{T})^2 \cdot L_\mathcal{H}}{c_\mathcal{H}^2 \cdot \ell_{all}}\right)$$

and assume that f has restricted smoothness $L_\mathcal{H}$ over $C_\mathcal{H}$ and restricted strong convexity ℓ_{all} over the sum of $C^\star$ with the negations of $C_\mathcal{T}$ and k copies of C, i.e., $\ell_{all} = \ell_{C^\star - C_\mathcal{T} - k \times C}$.

For a given $\varepsilon > 0$ and $R > \|\theta^\star\|_2$, 2PHASE-PGD *then returns an estimate $\hat{\theta}$ such that*

$$\|\hat{\theta} - \theta^\star\|_2^2 \;\leq\; \max\left(64 \cdot \frac{(1+c_\mathcal{T})^2}{\ell_{all}} \max_{\theta' \in (C^\star - C_\mathcal{T} - 1/c_\mathcal{H}^2 \times C) \cap \mathbb{S}^{d-1}} \langle \nabla f(\theta^\star), \theta' \rangle, \varepsilon R\right) \;.$$

The time complexity of 2PHASE-PGD *is dominated by $O(\log 1/\varepsilon)$ calls to the tail approximation and $O(k \log 1/\varepsilon)$ calls to the head approximation and the gradient oracle, respectively.*

[1] To avoid confusion: in this paper, $C^\star$ does not denote a polar cone, but instead the constraint set corresponding to the solution $\theta^\star$.

Since the theorem above is somewhat technical, we briefly instantiate it in the standard compressive sensing setting to illustrate the bounds. Due to space constraints, we unfortunately defer a more thorough discussion of the various parameters to the full version of the paper.

Let $C^\star$, $C_{\mathcal{H}}$, $C_{\mathcal{T}}$ each be the set of s-sparse vectors. Since we can project onto s-sparse vectors in linear time, we have $c_{\mathcal{H}} = c_{\mathcal{T}} = 1$. Hence our algorithm requires restricted smoothness and strong convexity only over sets that are $O(s)$-sparse, for which we can invoke standard results for RIP matrices [7]. The RIP setting also implies that $\ell_{all} \approx \ell_{\mathcal{H}} \approx 1$. Together, this yields the standard bound $\|\hat{\theta} - \theta^\star\|_2^2 \leq O(\|e\|_2)$ for a noise vector e after a sufficient number of iterations (so that the εR terms is negligible).

3 Algorithm

Algorithms 1 and 2 contain the two main parts of our new algorithm. As mentioned before, Algorithm 1 closely resembles the Frank Wolfe algorithm (conditional gradient method). Algorithm 2 is essentially projected gradient descent, but with INNERPHASE instead of a standard gradient step. We defer a more detailed discussion of these algorithms to the full version of the paper.

Algorithm 1 INNERPHASE

function INNERPHASE(θ^{in}, r, ρ)
  $k \leftarrow \left\lceil \frac{8L^{\mathcal{H}}}{\rho \cdot c_{\mathcal{H}}^2} \right\rceil$
  $\theta^{(0)} \leftarrow 0$
  for $i \leftarrow 0, \ldots, k-1$ **do**
    $g^{(i+1)} \leftarrow \text{HEADAPPROX}(-\nabla f(\theta^{in} + \theta^{(i)}))$
    $\tilde{\theta}^{(i+1)} \leftarrow \frac{r}{c_{\mathcal{H}}} \cdot \frac{g^{(i+1)}}{\|g^{(i+1)}\|_2}$
    $\theta^{(i+1)} \leftarrow \gamma^{(i)} \cdot \tilde{\theta}^{(i+1)} + (1 - \gamma^{(i)}) \cdot \theta^{(i)}$
  return $\theta^{in} + \theta^{(k)}$

Algorithm 2 2PHASE-PGD

function 2PHASE-PGD(R, ε)
  $T \leftarrow \left\lceil \log_2 \frac{1}{\varepsilon} \right\rceil$
  $\theta^{(0)} \leftarrow \vec{0}$
  $r^{(0)} \leftarrow R$
  for $t \leftarrow 0, \ldots, T-1$ **do**
    $\tilde{\theta}^{(t+1)} \leftarrow \text{INNERPHASE}(\theta^{(t)}, r^{(t)})$
    $\theta^{(t+1)} \leftarrow \text{TAILAPPROX}(\tilde{\theta}^{(t+1)})$
    if $\|\theta^{(t+1)} - \theta^{(t)}\|_2 > \frac{3}{2} r^{(t)}$ **then return** $\hat{\theta} \leftarrow \theta^{(t)}$
    $r^{(t+1)} \leftarrow \frac{r^{(t)}}{2}$
  return $\hat{\theta} \leftarrow \theta^{(T)}$

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 2012.

[2] Arturs Backurs, Piotr Indyk, and Ludwig Schmidt. Better approximations for tree sparsity in nearly-linear time. In *Symposium on Discrete Algorithms (SODA)*, 2017.

[3] Luca Baldassarre, Nirav Bhan, Volkan Cevher, Anastasios Kyrillidis, and Siddhartha Satpathi. Group-sparse model selection: Hardness and relaxations. *IEEE Transactions on Information Theory*, 2016.

[4] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010.

[5] Richard G. Baraniuk and Douglas L. Jones. A signal-dependent time-frequency representation: fast algorithm for optimal kernel design. *IEEE Transactions on Signal Processing*, 1994.

[6] Coralia Cartis and Andrew Thompson. An exact tree projection algorithm for wavelets. *Signal Processing Letters*, 2013.

[7] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[8] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Transactions on Information Theory*, 2015.

[9] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In *International Conference on Machine Learning (ICML)*, 2015.

[10] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. Fast recovery from a union of subspaces. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[11] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 2011.

[12] Prateek Jain, Nikhil Rao, and Inderjit S Dhillon. Structured sparse regression via greedy hard thresholding. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[13] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

[14] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time-data tradeoffs for linear inverse problems. *CoRR*, abs/1507.04793, 2015.

[15] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 2006.